모두의 말뭉치를 이용한 한국어 다의어 분별

신준철○, 이주상, 옥철영
울산대학교, 한국어처리연구실
ducksjc@gmail.com, dosa510@naver.com, okcy@ulsan.ac.kr

Korean Polysemy Word-Sense-Disambiguation using MoDu-Corpus

Joon-Choul Shin○, Ju-Sang Lee, Cheol-Young Ock
University of Ulsan, Korean Language Processing Lab

요 약

한국어 자연어처리 분야가 발달하면서 동형이의어 분별을 한 단계 넘어선 다의어 분별의 중요성이 점점 상승하고 있다. 최근에 다의어가 태깅된 "모두의 말뭉치"가 발표되었고, 이 말뭉치는 다의어가 태깅된 최초의 공개 말뭉치로써 다의어 연구가 본격적으로 진행될 수 있음을 의미한다. 본 논문에서는 이 말뭉치를 학습하여 작동하는 다의어 분별의 초기 모델을 제시하며, 이 모델의 실험 결과는 차후 연구를 위한 비교 기준점이 될 수 있다. 이 모델은 딥러닝을 사용하지 않은 통계형으로 개발되었고, 형태소분석과 동형이의어 분별은 기존의 UTagger로 해결하고 말뭉치 자원 외에도 UWordMap을 사용하여 다의어 분별을 보조하였다. 이 모델의 정확률은 약 87%이며, 다의어 분별 전에 형태소분석 또는 동형이의어 분별 단계에서 오류가 난 것을 포함한다. 현재까지 공개된 이 말뭉치는 오직 명사만 다의어 주석이 있기 때문에 명사만 정확률 측정 대상이 되었다. 이 연구를 통하여 다의어 분별의 어려움과, 다의어 분별에는 동형이의어 분별과는 다른 방법이 필요하다는 것을 확인할 수 있었다.

주제어: 모두의말뭉치, 다의어의미분별, 통계기반, UWordMap

1. 서론

자연어처리 연구와 응용에는 기본적으로 형태소분석이 기반기술로 필요로 하고, 나아가서는 동형이의어 분별 기술이 필요하였다. 현재에 와서는 더 높은 정확률의 기계번역이나, 질의응답 시스템을 위해 다의어 분별의 중요성이 상승하고 있다. 예를 들어서 사람의 신체부위 '손(Hand)'과 손님을 의미하는 '손(Guest)'이 있다. 이 둘은 어원이 다르며 동형이의어 수준에서 명확히 구분된다. 그러나 사람의 신체부위 '손(Hand)'과 일손에서 '손(Worker)'은 서로 어원이 같으며 동형이의어 수준에서는 구분할 수 없다. 따라서 동형이의어 분별만으로 이 둘을 구분할 수 없으며, 다의어 분별을 해야만 구분할 수 있다. 이 둘을 구분할 수 있게 되면 기계번역에서는 Hand와 Worker로 정확히 번역하는데 결정적인 도움이 되며, 검색이나 문장비교, 질의응답 등 다양한 자연어처리 분야에서도 큰 도움이 될 것이다.

다의어는 전문용어나 긴 단어에서는 찾아보기 힘들지만, 일상적으로 많이 사용하는 단어에서는 빈번하게 나타나며, 어원이 같음에도 그 의미가 크게 다를 수 있어 구분해야할 필요성이 있다. 예를 들어서 '살다'는 생명을 지니고 있다는 의미를 기본으로 하고 있으나 다양한 다의어를 가지고 있다. 다음의 예문들을 처리할 때 이를 구분한다면 수준 높은 자연어처리를 구현할 수 있다.

a-1 *그는 총을 맞아도 살 수 있다.*
a-2 *그는 백 살까지 살았다.*
b-1 *나는 아파트에 살고 있다.*
b-2 *우리 집에서 같이 살자.*
c-1 *그렇게 세게 부딪혔는 데도 시계가 살아 있다.*
c-2 *그 오래된 컴퓨터가 아직도 살아 있다.*

검색 분야에서 위 예문들이 모두 '살다'의 검색 결과로 나오기 보다는 다의어분별을 하여 a, b, c로 구분하여 검색되는 것이 더욱 편리할 것이다. 기계번역에서는 다의어 구분을 전처리 단계에서 미리 해둔다면 쉽게 정확한 대역어를 얻을 수 있다. a는 be alive나 survive 등으로 번역될 수 있고, b는 live, c는 work로 번역되는 식이다.

이렇게 다의어 분별이 큰 도움이 될 수 있음에도 다의어 분별 연구는 매우 적으며, 특히 한국어에서는 관련 연구를 찾아보기 힘든 수준이다. 거기에는 다양한 이유가 있지만 가장 큰 것은 2019년까지 공개된 대용량의 다의어 말뭉치가 없었기 때문이며, 최근에 국립국어원에서 모두의 말뭉치가 공개되면서 다의어 연구의 길이 열렸다. 본 논문은 이 말뭉치를 활용한 다의어 분별 모델과 그것을 실험한 내용을 제시한다.

한국어를 분석하여 다의어 분별까지 진행하기 위해서는 형태소분석과 동형이의어분별이 선행되어야 하는데, 본 논문에서는 이런 선행요소는 기존에 알려진 방법으로 채웠고. 다의어 분별 모델은 딥러닝이 아닌 통계형 방식으로 개발하였다. 이 다의어 분별 모델은 말뭉치 학습 정보와 함께 어휘지도를 동시에 사용하는데, 이를 위해 말뭉치기반과 지식기반의 하이브리드로 설계되었다.

2. 관련 연구

다의어 분별을 위해서는 우선 형태소분석과 동형이의어 분별이 되어야 하는데, 대표적으로 유태거(UTagger)가 있다[1, 2]. 유태거는 "기분석 부분 어절 사전"을 사용하여 어절마다 분석후보들을 생성하고, "부분어절

조건부확률" 을 이용하여 문맥에 맞는 후보를 선택한다. 동형이의어 분별까지 하였을 때 어절 단위 정확률은 96.5%이며, 속도가 매우 빠르고 사용자말뭉치 기술을 지원하고 있다. 본 논문에서는 유태거를 이용하여 동형이의어 분별을 하였고, 이후에 다의어 분별 모델이 적용된다.

류법모(2018)는 한국어 다의어 분별 기술을 활용하여 다국어 대역어 서비스를 개발하였다[3]. 이 서비스는 한국어를 입력받아 다의어 분별을 수행하고, 대역어를 찾기 위해 국립국어원에서 구축한 한국어기초사전 다국어 버전을 사용하였다. 여기에도 동형이의어 분별에는 유태거가 사용되었으며, 다의어 분별 정확률은 약 66%로 공개되어있다.

현재 한국어를 대상으로 가장 많은 정보를 가진 어휘의미망은 UWordMap(이하 UWM)이다[4]. UWM은 다의어 수준에서 구축되어 있기 때문에 이를 활용하면 지식기반 다의어 분별기를 만들 수 있는데, 배영준(2013)의 복합명사분석 연구가 있다[5]. 이 복합명사분석은 말뭉치에서 방향별 bigram 단위 학습데이터를 구축하고, 품사패턴과 UWM을 통해 학습을 확장하였다. 그 결과 86.20%의 정확률을 보였다. 신준철(2015)는 UWM만을 활용하여 명사와 용언 다의어를 대상으로 분별을 시도하였는데[6], 이 연구는 정확률 약 66%로 (명사 약 73%) 실용적인 수준이라고 보기는 힘들지만 다의어 분별에 UWM을 활용하는 방법을 제시하였다. 본 논문에서도 UWM을 활용하는 모듈이 있으며 신준철(2015)의 방법을 참조하였다.

신준철(2016)은 다양한 언어자원과 의미분별에 대해 연구하였고, 자원들 중에서 말뭉치가 가장 실용성이 좋고 어휘의미망이 보조적으로 사용될 수 있다고 확인하였다[7]. 비록 이 논문은 동형이의어 분별까지만 다루고 있지만, 다양한 자원들과 다양한 방법론을 논하고 있으며, 딥러닝 계열 연구인 Word Embedding을 의미분별에도 활용할 방법이 있음을 소개하고 있다. Word Embedding을 활용할 경우 말뭉치학습기반이 재현에 실패한 건에 대해 정확률을 다소 향상시킬 수 있을 것으로 보인다.

UWM을 이용한 동형이의어 분별 연구도 있다[8]. 이 연구에서는 명사-논항-용언으로 구성된 트리플 정보와 명사의 상하위 정보를 이용하여 동형이의어 분별 정확률을 향상시키는 방법을 제안하고 있으며, 의미 있는 수준의 정확률을 향상시켰다. 이 방법은 UWM을 다의어 수준에서 활용하면 다의어 분별에도 적용할 수 있으며, 본 논문에서 사용하는 모델이 이 방법을 참고하고 있으며 추가적으로 말뭉치 학습 방식도 사용하고 있다.

영어권에서는 다의어 분별을 위해 WordNet을 활용하는 연구가 있는데, Gemma Boleda(2012)는 WordNet을 활용하여 다의어 분별 모델 Centroid Attribute Model (CAM)을 개발하였다[9]. CAM은 단어를 표현하기 위해 단의어들을 사용하여 백터화 하고, 백터간 유사도를 구하여 다의어 분별을 한다. CAM은 실험 결과에서 특정한 경우에 최대 71%의 정확률을 보이기도 하였고 평균은 39.9%, 전체적으로 베이스라인을 넘는 성능을 보였다. Udaya Raj Dhungana(2017)는 영어 다의어 분별을 위해서 WordNet의 변형인 PolyWordNet을 사용하여 180개의 문장을 실험하였다[10]. 실험 규모가 작지만 비교적 최신 다의어 연구이며, WordNet의 단어는 다양한 다의어와 연결되어 있어 이것이 노이즈(Noise)를 일으키기 때문에 이를 해결하기 이 연구에서는 PolyWordNet을 구축하여 사용하였고, 의미있는 정확률 향상을 보였다.

3. 다의어 말뭉치와 의미번호체계

다의어 말뭉치가 있다면 다의어 분별 연구에 큰 도움이 되는데, 다의어 수준으로 주석을 다는 일은 상당히 고난도의 작업이여서 다의어 말뭉치 구축은 쉽지 않은 편이다. 최근에 국립국어원에서 모두의 말뭉치를 구축하였고 여기에 명사 한정으로 다의어 주석이 달린 말뭉치가 포함되어 있어 다의어 연구의 새로운 길을 열었다. 본 논문은 이 모두의 말뭉치를 활용하여 다의어 분별을 하는 방법을 보인다.

모두의 말뭉치에 있는 다의어 말뭉치는 문어와 구어로 분리되어 있으며, 문어는 2백만 어절, 구어는 1백만 어절로 구성되어 있고 정확한 규모는 표 1에 표시되어 있다. 기존에 널리 보급된 동형이의어 주석 말뭉치인 세종 말뭉치가 약 1천만 어절인 것을 생각해보면 이번에 공개된 다의어 말뭉치가 학습에 충분한 규모라고 보기는 힘들지만 실험은 충분히 가능하다고 볼 수 있다.

표 1 모두의 말뭉치 규모

	구어	문어
어절 수	1,006,447	2,000,213
문장 수	221,489	150,082
형태소 수	1,916,740	4,506,499

이 말뭉치는 명사에 한하여 다의어 주석이 달려 있으며, 따라서 본 논문에서도 실험은 명사에 대해서만 진행하였다. 명사 외의 어휘들에는 형태소분석과 품사 주석이 달려있다. 이 다의어 주석은 우리말샘 번호체계를 따르는데, 이 번호는 3자리로 표현되고 동형이의어 구분정보가 없다는 문제가 있다. 유태거를 이용하여 동형이의어 분별을 하게 되면 다의어 후보군을 줄일 수 있는데, 이런 방식을 사용하기 위해서는 다의어 번호에서 동형이의어 번호를 알 수 있어야 한다. 따라서 본 논문에서는 우리말샘 의미번호를 유태거가 사용하는 표준국어대사전2001 의미번호체계(이하 SDNS2001)로 변환하여 실험하였다[11]. 본 논문에서는 일반적으로 SDNS2001을 이용하여 서술하며, 우리말샘 의미번호인 경우에는 우리말샘 번호임을 표기한다. 우리말샘을 SDNS2001로 변환하는 과정에 일부 변환이 불가능한 사례가 있었는데, 표준국어대사전에 등록되어있지 않은 어휘들이며, 대체로 고유명사였다. 이런 어휘들은 동형이의어 번호를 알 수 없기 때문에 부득이하게 실험에서 제외되었다.

다의어 '유치' 는 뜻이 총 10개가 있으며, 우리말샘 방식으로는 3자리로 표시하고 SDNS2001로는 6자리로 표시하며 표 2에 자세히 표시하고 있다. 이 중에 010은 "행사나 사업 따위를 이끌어 들임." 이며, SDNS2001에서는 08_00_02에 해당한다. 처음 두 자리인 08은 동형이의어 구분 번호인데, 08에 속하는 다의어는 08_00_01과 08_00_02로 2개뿐이다. 이런 경우에 동형이의어 분별 정

보를 이용한다면 후보를 10개에서 2개로 줄일 수 있게 된다.

표 2 우리말샘과 표준국어대사전 의미번호 비교

우리말샘	표준국어대사전2001
유치001	유치01_00_00
유치002	유치02_00_00
유치003	유치03_00_00
유치004	유치04_00_00
유치005	유치05_00_00
유치006	유치06_00_00
유치007	유치07_00_01
유치008	유치07_00_02
유치009	유치08_00_01
유치010	유치08_00_02

자연어처리로 다의어 자동분별을 하기 위해서 우리말샘보다 SDNS2001의 방식이 더 용이한데, 동형이의어 분별과 다의어 분별을 분리하여 처리할 수 있으며, 동형이의어 분별까지에 기존의 연구를 사용할 수 있기 때문이다. SDNS2001은 우리말샘 번호체계에 비하여 의미번호의 길이가 3자리 길어지지만, 이 외의 단점은 없다.

4. 시스템

4.1. 전체 구성

본 논문에서 실험에 사용한 시스템은 여러 모듈로 구성되어지는데, 처음에는 사용자가 입력한 한국어 원시 문장을 유태거가 형태소분석과 동형이의어 분별을 한다. 이렇게 분석된 데이터에는 동형이의어 분별 정보가 포함되어 있는데, 이 시스템은 동형이의어에 속하는 다의어 후보들을 표준국어대사전과 말뭉치학습사전을 참조하여 생성한다. 예를 들어 '유치하였다'가 입력되면 이를 유태거가 분석하여 "유치08NNG+하..." 같은 형태로 만든다. 그 다음으로 동형이의어 번호 정보를 통해서 다의어후보 정보를 생성하는데, '유치08_00_01(꾀어서 데려옴.)'과 '유치08_00_02(행사나 사업 따위를 이끌어 들임.)' 두 개의 뜻을 가지고 있기에 2개의 후보가 만들어 진다. 다음으로 이 데이터는 후보 중에서 문맥상 적합한 것을 고르는 "다의어 분별" 모듈을 거치게 되고, 최종적으로 시스템은 다의어 분별 결과를 출력한다. 전 과정은 그림 1에서 순서도로 표현하고 있다.

4.2. 말뭉치기반 다의어 분별

본 모델에서는 어절 경계 정보를 사용하지 않고, 문장을 형태소들의 나열로만 취급한다. 예를 들어 문장 "대회 본선 무대를 밟았으며"는 다음의 형태로 표현될 수 있다.

대회02_00_02NNG 본선03_00_00NNG 무대06_00_02NNG 를JKO 밟VV 았EP 으며EC

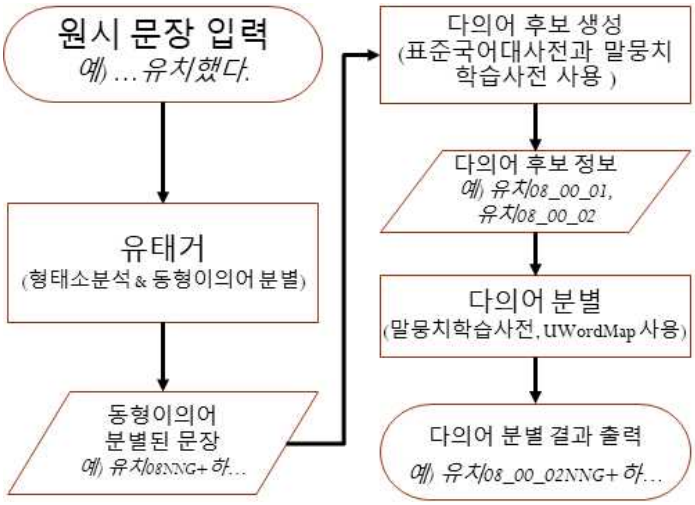

그림 1 시스템 순서도

형태소를 동형이의어 수준까지 표현한 것을 h라고 정의하고, 다의어는 poly로 정의한다. 예로 무대060002NNG는 poly이며 이것의 h는 무대06NNG이다. 다의어 분별 모델의 관점에서 입력 정보는 이미 동형이의어 분별이 끝난 문장이기 때문에 h의 나열로 정의할 수 있고, 각 h는 후보들을 poly의 집합 형태로 가진다. 이를 식 (1)에서 자세히 표현하고 있다. 상술한 예의 문장에서 h_1은 '대회02NNG'이고 h_2은 '본선03NNG'이다. $poly_i$는 i번째 위치한 형태소의 다의어 후보 집합이다. 본 알고리즘의 최종 목적은 저 집합에서 가장 적절한 1개를 선택하는 것이다.

본 모델은 다수의 자질 함수를 사용하여 다의어 분별을 수행하는데 그 중 첫 번째인 f1은 인접 형태소를 전혀 고려하지 않은 확률을 계산한다. 만약 학습말뭉치에 '세계02NNG'가 1000번 등장하고 그 중에서 '세계020005NNG'가 900개라면, f1(세계020005NNG)는 학습말뭉치를 기반으로 계산하여 0.9를 반환한다.

$$
\begin{aligned}
&Sent = h_1, h_2, h_3 \ldots h_n \\
&poly_i = \{poly_{i,1}, poly_{i,2} \ldots poly_{i,m}\} \\
&f1(poly_{i,j}) = p(poly_{i,j} | h_i)
\end{aligned}
\quad (1)
$$

일반적으로 '무대06NNG'는 주로 무대06_00_01NNG(노래, 춤, 연극 따위를 하기 위하여 객석 정면에 만들어 놓은 단.)으로 사용되는데, '본선' 다음에 위치한 경우에는 무대06_00_02NNG(재능, 솜씨 따위를 나타낼 수 있게 된 판)의 의미를 가진다. 이런 경우에 다의어 분별을 적절히 수행하기 위해서는 '본선'처럼 인접한 형태소를 이용해야하고, 따라서 윈도우2에 해당하는 자질 함수를 사용해야한다. 현재 대상 형태소 (예:무대) 기준에서 좌측의 형태소를 (예:본선) 이용하는 것을 f21이라 하고 우측의 형태소를 (예:를JKO) 이용하는 것을 f2r이라 정의한다. 이 자질함수는 식 (2)에서 자세하게 표현하고 있다.

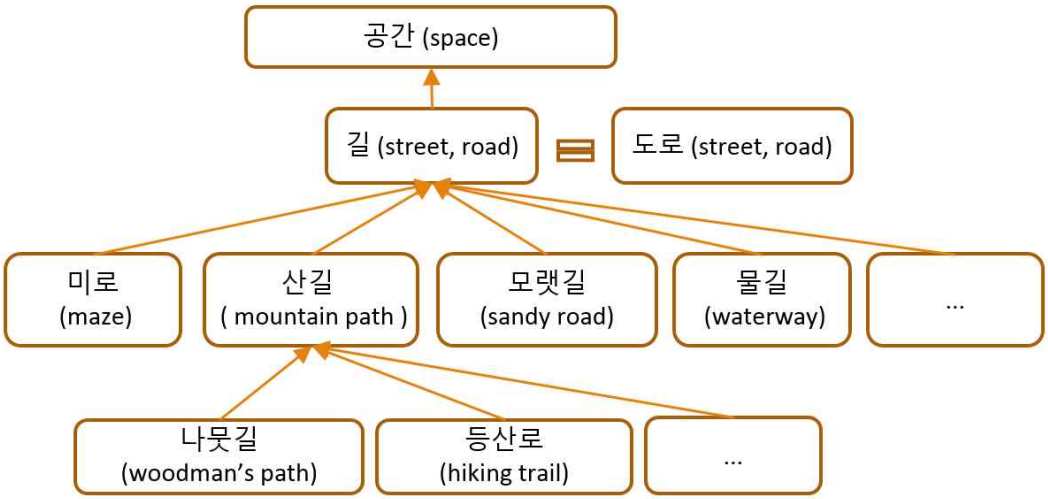

그림 2 UWordMap에서 '길(street)'

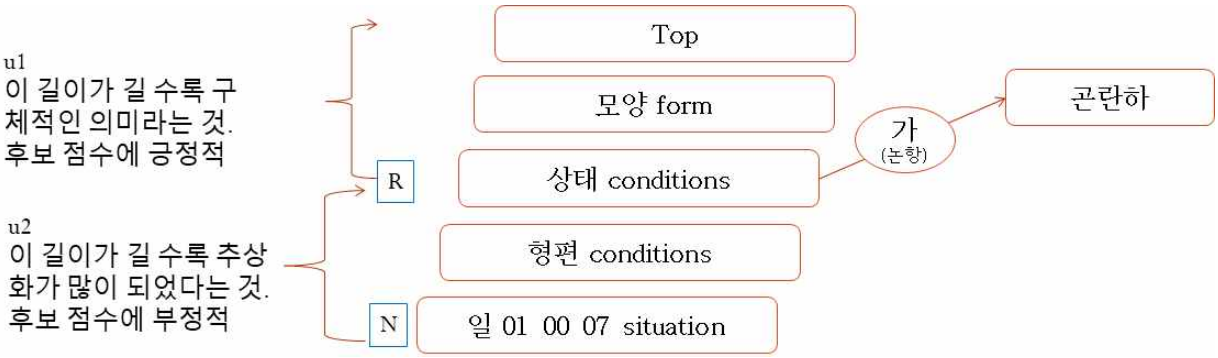

그림 3 상위어와 하위범주화 정보

$$f2l(poly_{i,j}) = p(poly_{i,j}|h_{i-1}, h_i)$$
$$f2r(poly_{i,j}) = p(poly_{i,j}|h_i, h_{i+1}) \quad (2)$$

만약 예문에서 '무대'의 다의어 분별을 위해 '대회'도 사용할 수 있다면 더욱 정확하게 분별할 수 있을 것인데, 이런 식으로 윈도우 크기를 늘려가면 정확률을 향상되겠지만 전체 알고리즘이 크고 무거워지며 윈도우가 큰 자질 함수는 재현율이 낮아져서 실용성이 내려간다. 따라서 본 논문에서는 윈도우 4까지 사용하는 것을 제안한다.

윈도우가 큰 함수일수록 긴 형태소열을 이용하기 때문에 빈도만 충분하다면 더욱 중요한 정보이다. 따라서 가중치를 주어서 차별화할 필요가 있다. 최종적으로 f들과 각 f에 대한 가중치 w의 "곱의 합"을 계산하며, 이것은 수식 (5)의 형태가 된다.

$$Tag(h_i) = \text{arg}max_j \sum_n f_n(poly_{i,j}) \times w_n \quad (5)$$

4.3. UWordMap기반 다의어 분별

상술한 말뭉치기반 모델은 학습말뭉치에 나타난 페턴에 대해서만 재현이 가능하다. 재현율을 올리기 위해서는 말뭉치를 늘리거나, 말뭉치 회의 학습자원을 사용해야 한다. UWM에 있는 명사 상하위 관계 정보와, 명사-논항-용언의 정보로 구성된 하위범주화를 이용한다면 아주 다양한 패턴들을 처리할 수 있다.

그림 2는 UWM에 등록된 '길(street)'에 대한 정보다. '길'에는 산길이나 모랫길 등 다양한 길이 있으며, 이것들은 용언 '걷다'와 같이 사용하면 자연스러운 문장을 구성할 수 있다. 현실적으로 모든 '길'의 하위어들이 '걷다'와 같이 등장하도록 말뭉치를 구성하는 것은 어렵고 비효율 적이며, UWM에서 '길'의 하위어 정보와 하위범주화 정보를 조합하여 이 문제를 해결할 수 있다.

예를 들어서 "등산로를 걷다"를 분석할 때 '등산로'와 '걷다' 간의 하위범주화 관계가 있는지 확인을 해보게 되고, 관계가 없다면 '등산로'의 상위어인 '산길'과 '걷다'가 관계있는지 확인해보게 된다. 이를 반복하여 최상위어까지 올라가면서 하위범주화 정보를 확인한다. 반대방향으로 하위어를 확인해볼 수도 있는데, 한 단어의 하위어를 계속 탐색하게 되면, 탐색해야할 단어의 수가 굉장히 많아질 수 있기 때문에 현실적으로 속도의 문제가 발생할 수 있어서 이 모델에서는 상

위어만 탐색하는 방법을 사용하였다.

상위어는 항상 1개만 존재하며, 상위어를 추적하는 과정에서 처음으로 하위범주화 정보를 발견한 단어를 R이라고 정의한다. 그리고 현재의 단어를 N이라고 정의한다. 예를 들어 “그 일은 곤란하다.”에서 ‘일’의 다의어 분별을 수행할 때, ‘일’은 ‘곤란하다’와 직접적인 하위범주화 관계에 있지는 않으며, ‘일’의 2단계 위 상위어 ‘상태’에서 ‘곤란하다’와 관계를 가진다. 이 예에서 현재 단어 N은 ‘일’이고 하위범주화 관계를 찾은 상위어는 R이다. 이 예는 그림 3에서 표현하고 있다. 일반적으로 명사가 최상위어(Top)와 멀리 있을수록 더욱 구체적인 명사라고 표현할 수 있고, 그런 명사와 직접적으로 연관된 하위범주화 정보가 확인되었다는 것은 이 관계정보에 대해 더 신뢰할 수 있다는 의미이다.

N과 R이 멀리 떨어질수록 하위범주화 관계에 있는 그 용언을 N에 사용하기에 부적절할 확률이 증가한다. 예를 들어서 그림 3를 보면 확인된 하위범주화 정보는 상태-가-곤란하다 이다. “형편이 곤란하다”는 여전히 자연스러운 표현이며, “일이 곤란하다” 까지는 어색하지 않다. 그러나 일01_00_07의 하위어로 예를 들어 ‘경사(축하할 만한 기쁜 일)’에 ‘곤란하다’를 같이 사용하면 조금 어색해진다. 따라서 N과 R은 가까울수록 다의어 N이 정답일 확률이 높다고 추측된다. 그래서 R과 N사이의 거리가 클수록 작은 값이 나오도록 함수를 구성하면 다의어 태깅에 도움이 될 것이다. 이를 위해 거리에 1 더한 값의 역수를 사용할 수 있으며, 이것을 u2라고 정의한다.

최상위어에 도달할 때 까지 아무런 하위범주화 정보를 확인하지 못할 수도 있다. 이런 경우에 해당 N은 정답이 아닐 확률이 매우 높다. 따라서 하위범주화 정보가 존재할 경우 1을 반환하고, 아니면 0을 반환하는 함수 u3을 정의한다. 표 3에 모든 UWM 자질함수를 정리하였다. 전체 시스템은 이 자질함수들도 식 (5)에 포함하여 다의어 분별을 수행한다.

표 3 UWordMap 자질 함수

자질 함수 이름	반환 값 계산
fu1	최상위어에서 R까지 거리
fu2	1/((N에서 R까지의 거리)+1)
fu3	하위범주화 파악 여부 파악 됨 = 1 없음 = 0

5. 실험 결과

학습을 위해서 울산대학교에서 구축한 표준국어대사전 다의어 말뭉치 전체를 사용하였고, 모두의 말뭉치 문어와 구어에서 80%를 학습하였다. 나머지 20%는 학습하지 않고 정확률 측정용으로 사용하였고, 정확률 측정 대상은 다의어 주석이 되어 있는 명사류에서 오직 NNG(일반명사)만 포함하였으며, 모두의 말뭉치에 ‘~하다’와 같은 어간형 형태소들은 모두 어근형(명사+하다)으로 분석되어 있기 때문에 이런 명사들도 정확률 측정에 포함되었다. 다만 일부 새로운 명사들은 그 의미번호가 SDNS2001로 표현되지 못하여 부득이하게 제외되었다. 동형이의어는 맞으나 다의어는 아닌 형태소도 제외되었다. 예를 들어 ‘본선03NNG’는 다의어 번호표기법으로 표현할 경우 ‘본선03_00_00NNG’이며, 이는 ‘본선’을 동형이의어 분별하여 ‘본선03’까지 확정할 경우에 더 이상 분별할 다의어가 없다는 것을 의미한다. 실험 결과 정확률은 평균 86.91%가 나왔으며, 문어와 구어 개별 정확률과 자세한 내용은 표 4가 나타낸다.

표 4 다의어 분별 정확률

말뭉치	정확률	정답 수	측정 대상
문어	87.63%	113,144	129,115
구어	84.39%	31,095	36,846
전체	86.91%	144,239	165,961
UWM2015 명사[6]	72.93%	339,515	465,526

신준철(2015)[6]와 비교를 위해 표 4에 UWM2015로 표기된 정보가 있다. 이 연구에도 명사만 분리하여 정확률을 측정한 내용이 있으며 그 정확률이 본 논문에서 제안하는 방법보다 약 14% 포인트 낮은 것을 볼 수 있다. 비록 말뭉치가 다르긴 하지만 정확률에 상당한 차이가 나며, 이것은 UWM만 사용하는 것 보다 말뭉치를 학습하는 것이 정확률 향상에 큰 도움이 되기 때문인 것으로 분석된다.

오답에는 형태소분석이나 품사 분별 또는 동형이의어 분별에서 틀린 것이 포함되어 있으며, 동형이의어 분별까지 사용된 유태거의 알려진 정확률은 96.5%이다. 즉, 정확률 측정 대상 100개 당 3~4개는 다의어 분별 전에 오류가 발생한 것으로 추정되며, 10개 정도가 다의어 분별에 실패한 것으로 볼 수 있다. 예를 들어서 ‘이밖에’가 “이MMD 밖00_00_03NNG 에JKB”에로 정답이 등록되어 있는데 유태거에서 “이NP 밖에JX”로 분석하여 오답으로 처리되었다. 전체 문맥으로 판단하면 모두의 말뭉치가 형태소분석을 잘못한 경우이다. 이런 식으로 정확률 실험에는 다의어 분별 전에 오답이 나온 경우도 포함되어 있고, 말뭉치의 오류도 포함되어 있다.

문어의 정확률이 약 3% 포인트 더 높게 나왔는데, 이는 일반적으로 형태소 원형 복원이나 품사 구분, 동형이의어 분별 등 전반적인 과정에서 구어보다 문어의 정확률이 높기 때문인 것으로 판단된다. 또한 구어의 말뭉치 규모가 문어에 비해 절반이며, 유태거 학습말뭉치에서도 구어의 규모가 문어보다 훨씬 적은 편이다.

세종말뭉치와 달리 모두의 말뭉치는 현재 시점에서 명사만 다의어 주석되어 있고, 나머지 품사들은 동형이의어 분별도 되어있지 않다. 예를 들어 “차를 타서 마셨다”를 분석할 때 ‘타’의 동형이의어 정보가 없다면 ‘차’를 분석하기가 쉽지 않다. 이런 문제 때문에 일부 오답이 발생한 것으로 추측된다.

UWordMap의 다의어 사전과 우리말샘의 차이로 인한 오류도 존재한다. 우리말샘에는 ‘일부02NNG’의 뜻이 “한 부분. 또는 전체를 여럿으로 나눈 얼마.”로 등록되어 있는데, UWordMap에 등록하는 과정에서 이 단어의

상위어를 ‘부분’과 ‘얼마’로 나누기 위해 ‘일부02_00_01’과 ‘일부02_00_02’를 만들게 되었다. UWordMap은 모든 단어가 상위어를 가지며 반드시 1개만 가진다는 규칙이 있기 때문이다. 이런 불일치로 인해 일부 오류가 발생하였다.

6. 결론

본 논문은 최근에 공개된 “모두의 말뭉치” 중 어휘의 미분석말뭉치를 이용한 다의어분별 시스템을 제안하였다. 실험에서 정확률 측정 대상은 약 17만개로 충분하다고 할 수도 있지만, 약 3백만 어절의 말뭉치 규모는 다의어 분별을 위한 학습용으로 충분하다고 보기는 어렵다. 그러나 본 논문에서 실험한 모델은 약 87%의 정확률을 보여주고 있으며, 일부 분야에서는 이정도 정확률로도 실용성이 있을 것으로 기대되며, 이후에 연구될 다의어 분별 분야에서 기준선이 될 수 있을 것으로 판단된다. UWordMap과 우리말샘의 차이, 그리고 모두의 말뭉치 내의 오류 등으로 인해 정확률이 다소 떨어지는 것을 확인하였으며, 이러한 문제점들이 수정된다면 정확률이 크게 향상될 것으로 보인다. 또한 이 시스템은 딥러닝을 사용하지 않았기 때문에 학습 및 처리가 빠르며 분석 과정을 추적할 수 있어 앞으로 정확률을 향상시킬 여지가 많다.

이 시스템은 모두의 말뭉치뿐만 아니라 표준국어대사전 말뭉치와 UWordMap도 같이 활용하고 있으나, 각 모델이 다의어 분별에 미치는 영향에 대한 분석이 되어있지 않으며, 이러한 내용을 차후에 연구할 필요가 있다. 그리고 모두의 말뭉치에는 NNG(일반명사) 외에도 다른 고유명사, 대명사들의 다의어 주석 정보를 포함하고 있기 때문에 이런 부분에 대해서도 다의어분별을 할 수 있게 모델을 개선하고 실험할 필요가 있다. 특히 2021년에 다의어 용언의 말뭉치도 공개된다면 의존관계에 있는 용언과 명사와의 다의어 분별 실험이 가능해질 것이기에 이런 연구들이 앞으로 진행될 필요가 있다.

현재 모두의 말뭉치에는 우리말샘 번호체계가 적용되어 있는데, 이로 인하여 동형이의어 번호를 알 수 없는 문제가 있다. 본 연구진은 이번 시스템을 개발하고 실험을 하는 전 과정에서 이 의미번호체계의 문제로 인해 많은 어려움이 있었으며, 번호 차이로 이하여 부득이하게 실험에서 제외된 부분도 존재한다. 자연어처리 관점에서는 동형이의어 번호를 즉시 구분할 수 있는 표준국어대사전의 번호체계가 더욱 용이하다는 것을 확인할 수 있었다. 우리말샘 사이트에는 동형이의어 수준에서 같은 의미의 다의어들은 같이 묶여서 출력되고 있는데, 그렇다면 동형이의어 정보가 존재하다는 것이고 그런 사전 정보들이 말뭉치와 함께 제공될 필요가 있겠다.

감사의 글

이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단-리서치펠로우의 지원(NRF-2017R1A6A3A11034211, 다의어 분별과 사용자 말뭉치 연구)과 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2013-0-00131, (엑소브레인-총괄/1세부) 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발)의 지원을 받아 수행된 연구임.

참고문헌

[1] Joon-Choul Shin, C. Y. Ock, "A Korean Morphological Analyzer using a Pre-analyzed Partial Word-phrase Dictionary", Journal of KIISE : Software and Applications, Vol. 39, No. 5, pp. 415-424, May. 2012. (in Korean)

[2] Joon-Choul Shin, C. Y. Ock, "Korean Homograph Tagging model based on Sub-Word Conditional Probability", Journal of KIPS : Software and Data Enginering, Vol. 3, No. 10, pp. 407-420, Oct. 2014. (in Korean)

[3] Pum-Mo Ruy, "Multilingual Word Translation Service based on Word Semantic Analysis ", Journal of Digital Contests Society, Vol. 19, No. 1, pp. 75-83, Jan. 2018. (in Korean)

[4] Young-jun Bae, C. Y. Ock, "Intorduction to the Korean WordMap(UWordMap) and API", Proc. of 26th Annual Conference on Hamman and Cognitive Language Technology, Vol. 26, No. 1, pp. 27- 31, 2014. (in Korean)

[5] Young-Jun Bae, "Semantic Analysis of Korean Compound Noun using Lexical Semantic Network(U-WIN)", Journal of KIISE, Vol. 40, No. 12, pp. 833-847, Oct. 2013. (in Korean)

[6] Joon-Choul Shin, C. Y Ock, “Noun and Verb Polysemy Word Sense Disambiguation Using UWordMap”, Proc. of 27th Annual Conference on Human & Cognitive Language Technology, pp. 216-219, Oct. 2015.

[7] Joon-Choul Shin, C. Y Ock, “Semantic Resources for Korean Semantic Analysis and Word Sense Disambiguation”, Journal of KIISE, Vol. 34, No. 8, pp. 8-16, Aug. 2016.

[8] Joon-Choul Shin, C. Y Ock, “Improvement of Korean Homograph Disambiguation using Korean Lexical Semantic Network (UWordMap)”, Journal of KIISE, Vol. 43, No. 1, pp.71-79, Jan. 2016. (in Korean)

[9] Gemma Boleda, Sebastian Pado, Jason Utt, "Regular polysemy: A distributional model ", Proc. of SEM Conference, pp. 151-160, Jun. 2012.

[10] Uraj Dhungana, "PolyWordNet: A Word Sense Disambiguation Specific WordNet of Polysemy Words", Dissertation, Institute of Engineering Trihuvan University, Aug. 2016.

[11] 우리말샘-표준국어대사전(2001) 연결 자료, https://www.korean.go.kr/front/etcData/etcDataView.do?mn_id=208&etc_seq=642, Dec. 2019.